Causality

… a brief overview

Jason Hartford


Why should I care?
Margaret Ellis
THE ROAD NOT TAKEN

Why should I care?

Practical

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren’t pure prediction problems. They care about designing policies (interventions)

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)

- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)
- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.
- Can we repurpose some of the tools we've built for this data for causal inference?

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)
- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.
- Can we repurpose some of the tools we've built for this data for causal inference?

Ambitious AI goals

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)
- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.
- Can we repurpose some of the tools we've built for this data for causal inference?

Ambitious AI goals

- One motivation for unsupervised learning is: let's find ways to model the world so that we can plan in the model before we interact in the real world ("imagine" what might happen).

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)
- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.
- Can we repurpose some of the tools we've built for this data for causal inference?

Ambitious AI goals

- One motivation for unsupervised learning is: let's find ways to model the world so that we can plan in the model before we interact in the real world ("imagine" what might happen).
- If we could learn $\hat{p}(x, y) \approx p(x, y)$ from observing the world - maybe we could plan: $x^* \approx \text{argmax}_x \hat{p}(y \mid x)$.

Why should I care?

Practical

- Most questions in social science, medicine, etc. aren't pure prediction problems. They care about designing policies (interventions)
- They have rich, high dimensional data (e.g. text, images, etc.) but no good methods for dealing with it.
- Can we repurpose some of the tools we've built for this data for causal inference?

Ambitious AI goals

- One motivation for unsupervised learning is: let's find ways to model the world so that we can plan in the model before we interact in the real world ("imagine" what might happen).
- If we could learn $\hat{p}(x, y) \approx p(x, y)$ from observing the world - maybe we could plan: $x^* \approx \text{argmax}_x \hat{p}(y \mid x)$.
- Problem: this violate IID assumption. Causal inference gives concrete cases when this is possible and when it isn't.


Margaret Ellis
THE ROAD NOT TAKEN

The Road Not Taken by Robert Frost

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Potential outcomes… two roads

- “**Treatment**”, T, is a dependent variable you care about (‘which road?’), “**response**”, Y, is some outcome of interest (‘life happiness’) and X are (potentially confounding) features / context.

- For the next couple of slides, let’s assume a binary treatment $T \in \{0,1\}$. Each person (‘**unit**’) has two roads that they could go down (‘**potential outcomes**’ / ‘**factual** and **counterfactual**’ outcomes). Call these Y(1) and Y(0). You only ever observe one of the two outcomes.

- Simplest question we’d like to ask: did *“[taking] the road less traveled”* make all the difference? What is $Y_i(1) - Y_i(0)$? How about $\mathbb{E}[Y_i(1) - Y_i(0)]$?

Isn’t this just supervised learning?

Isn’t this just supervised learning?

- Give me a bunch of labeled examples of people who took the left road and people who took the right road and I’ll fit you a model that gives you $\mathbb{E}[Y \mid T = t]$.

Isn’t this just supervised learning?

- Give me a bunch of labeled examples of people who took the left road and people who took the right road and I’ll fit you a model that gives you $\mathbb{E}[Y \mid T = t]$.

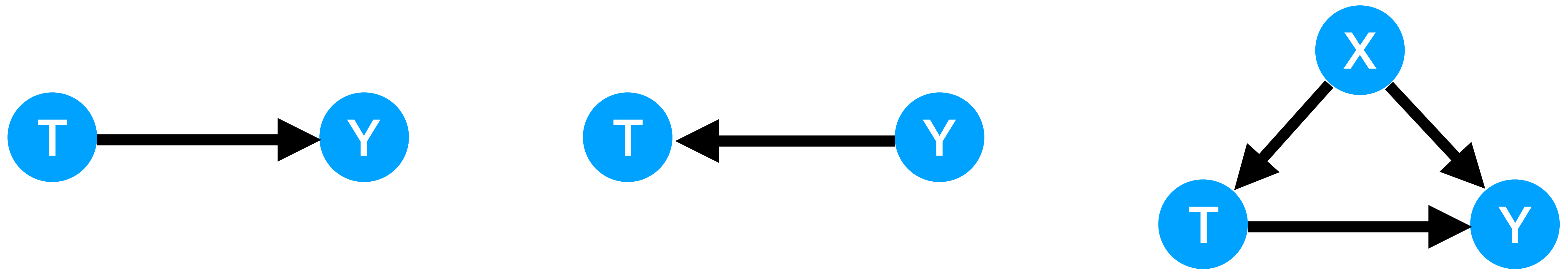
- What can go wrong with this strategy?

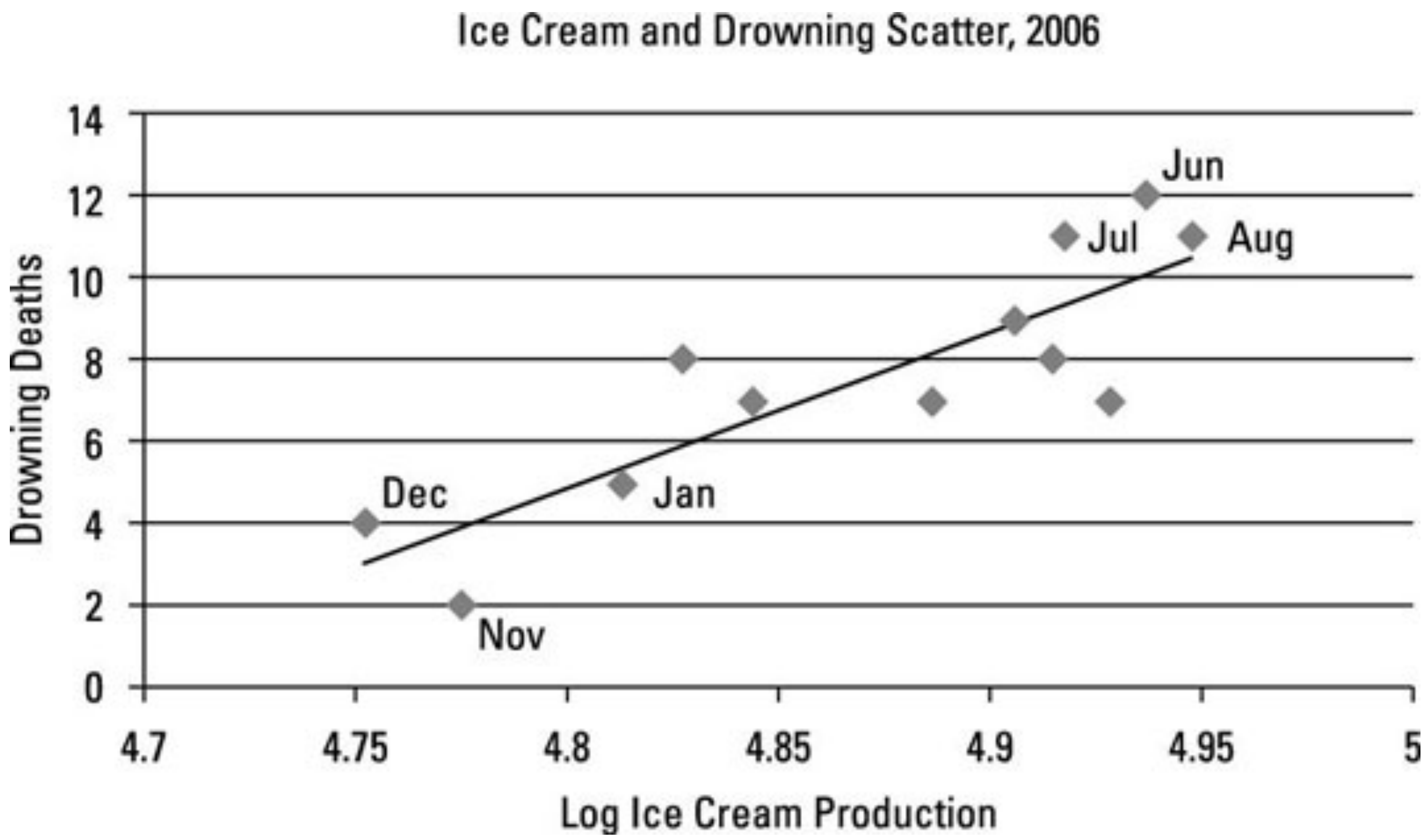
Isn't this just supervised learning?

- Give me a bunch of labeled examples of people who took the left road and people who took the right road and I'll fit you a model that gives you $\mathbb{E}[Y \mid T = t]$.

- What can go wrong with this strategy?

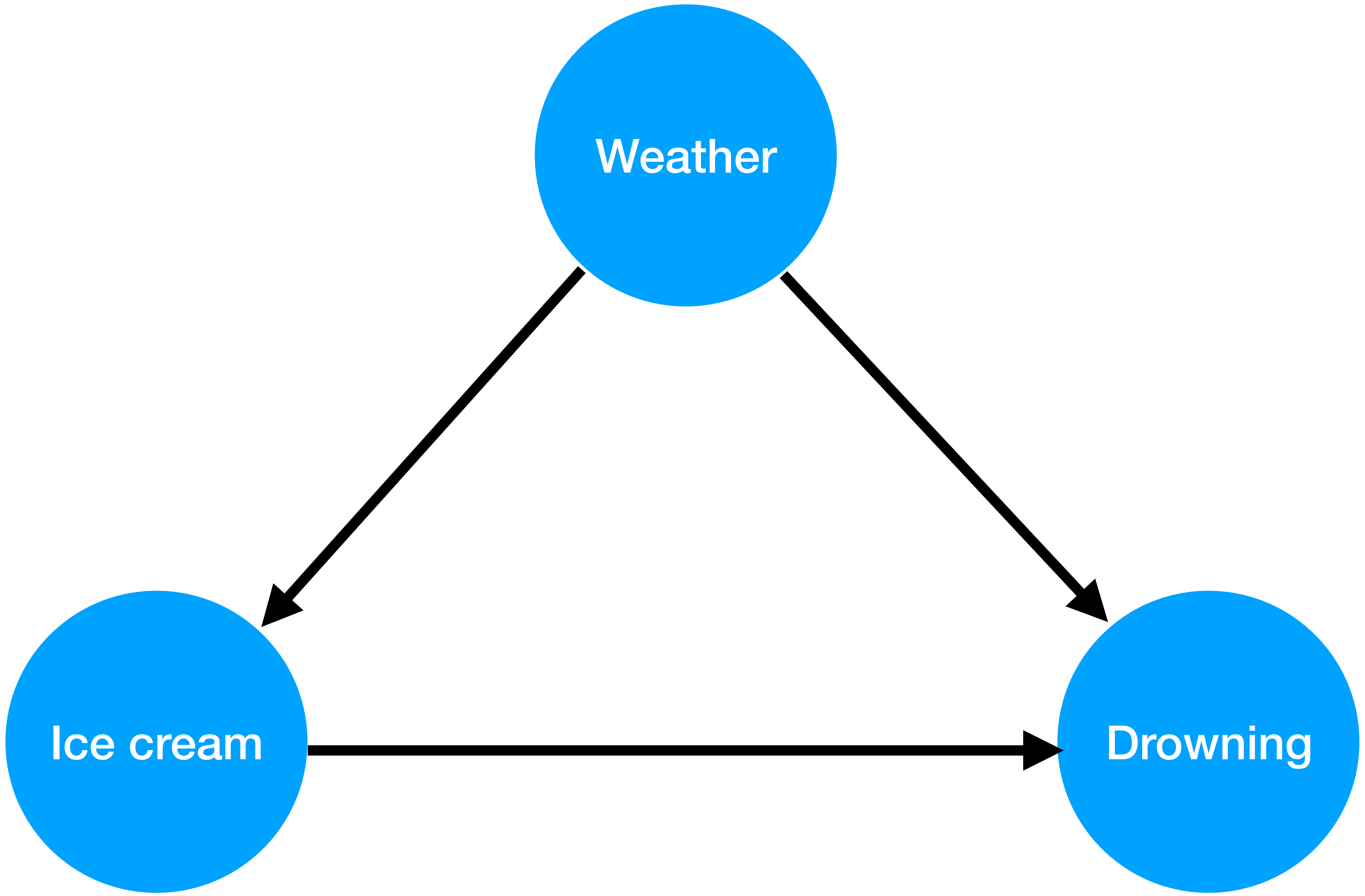
- Correlation $\neq$ causation. If two variables are correlated we may be in one of three scenarios:





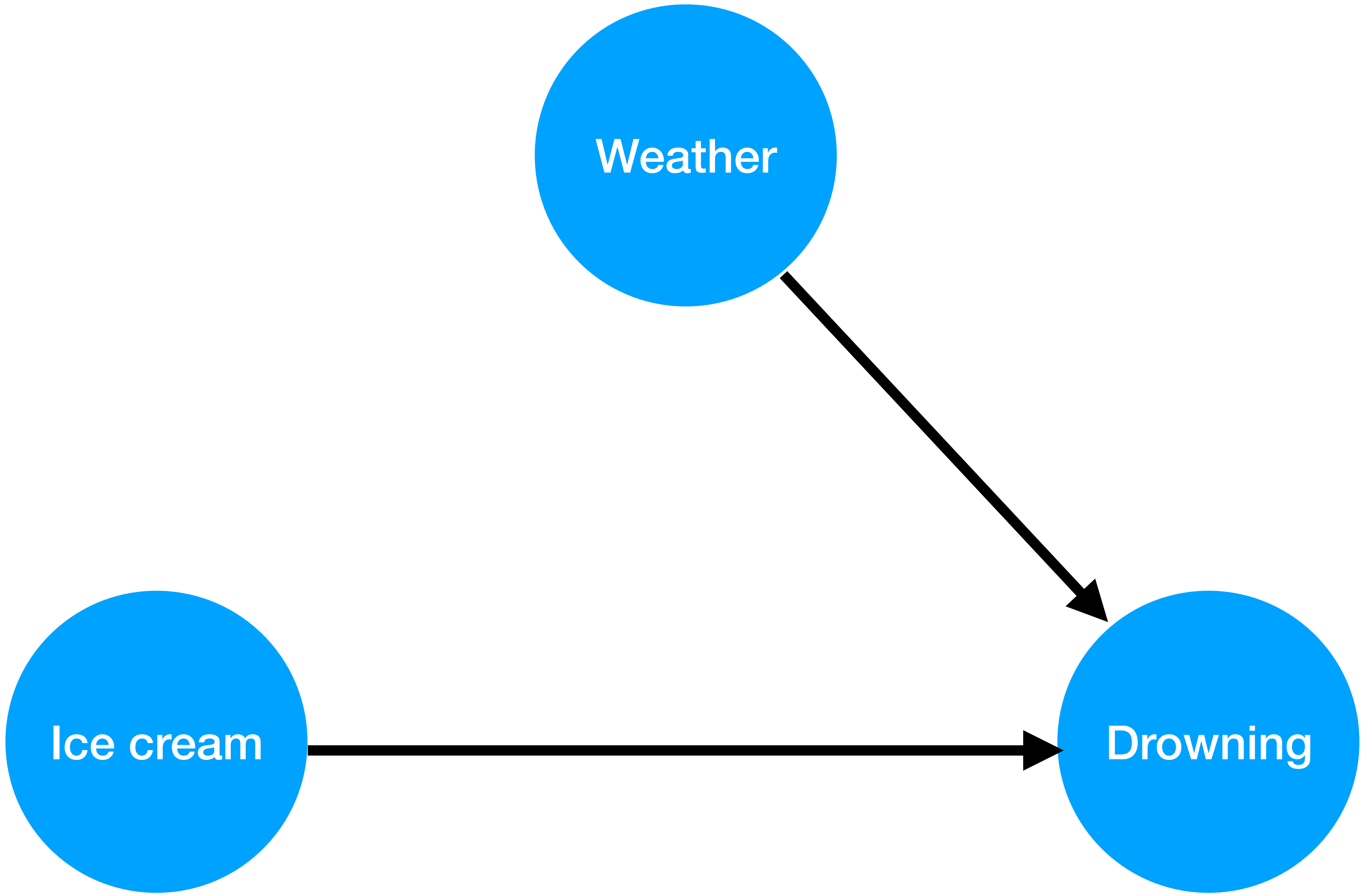
Stolen from Victor Veitch's slides / Econometrics for dummies

What went wrong?



Supervised learning predicts $E[y \mid t]$. That will do a good job of predicting under the observational distribution.

What went wrong?



Supervised learning predicts $E[y \mid t]$. That will do a good job of predicting under the observational distribution.

But if we want to know if we should ban ice cream, we need to know about $E[y \mid \mathrm{do}(t)]$, which is a different distribution.

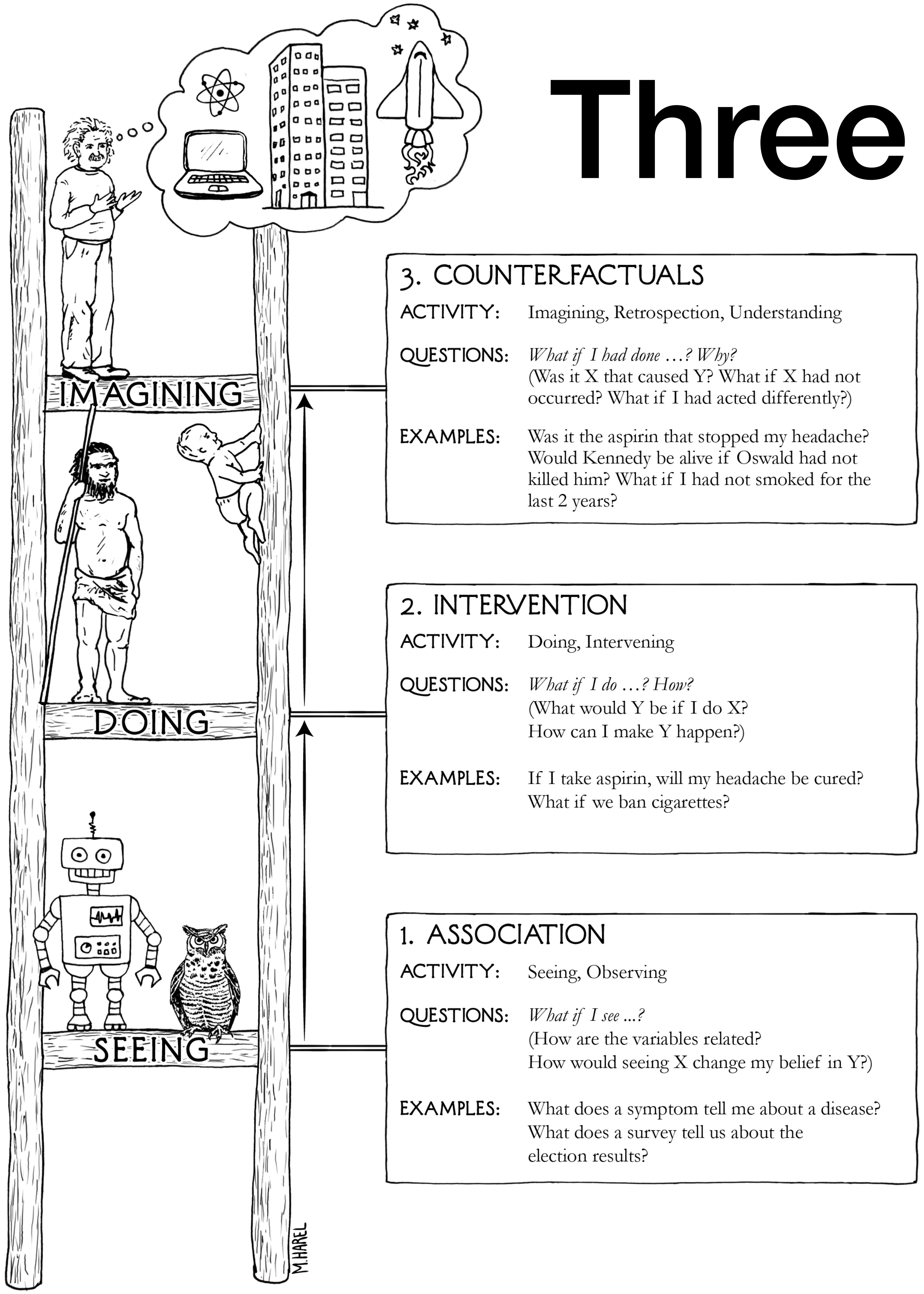
How do we solve it?

- Option 1: **Randomized control trails** (A/B testing / online learning). Collect data that explicitly randomizes over the treatment and measures the response. So: $p(y, t) = p(y, \text{do}(t))$.

- Option 2: Estimate the $\mathbb{E}[y \mid t, x]$ for each temperature x. Any remaining effect must be the result of t if there are no additional confounders. '**Backdoor adjustment**' formula:

- $\mathbb{E}[y \mid \text{do}(1)] - \mathbb{E}[y \mid \text{do}(0)] = \mathbb{E}_x \left[\mathbb{E}[y \mid t = 1, x] - \mathbb{E}[y \mid t = 0, x] \right]$

Three levels of questions



Observational questions

Do people who are given the drug tend to recover?

Action/Intervention Questions

If I give people this drug, how likely it is that they recover?

Counterfactuals

The patient survived. Had I not given the patient the drug two weeks ago, would she still have recovered?

What will we cover this block?

- The simple backdoor adjustment formula idea can be generalized to more complex graphs. Some key ideas to solve them - **backdoor criterion**, **do calculus** and **front door adjustment**. Estimation with deep nets.

- All of these methods only work when we can “block” all confounding effects. What happens when we have **unobserved** confounders? **Instrumental variable** methods and (in some cases) **proxy** variables.

- What if we don’t have the graph? Can we learn it from data? **Causal discovery** studies this….

What will we cover this block?

- At its core, causality is about generalizing from $p(y, x)$ to $p(y, \mathrm{do}(x))$. Are there other ways we can generalize beyond $p(y, x)$? **Invariant risk minimization** studies this from a representation learning perspective.

- Extensions to **causal bandits**, **reinforcement learning** and **causal inference on images**.

- Pearl's notion of **counterfactuals** - what you can do with them and what makes them hard.

- Bengio et al.'s attempts at **causal discovery** via learning.

What we won’t cover

Practical

- Sensitivity analysis
- Doublely robust estimators
- “Double Machine Learning”
- Most of causal discovery
- Data fusion

Open questions

- Causal inference on text and images (in its full generality).
- Causal models of environments for model-based RL.
- etc.